Expert assessment vs. machine learning algorithms: juvenile criminal recidivism in Catalonia

Songül Tolan (JRC), Carlos Castillo (UPF), Marius Miron (JRC), Emilia Gómez (JRC)

Algorithms & Society Workshop, Brussels, 10 December 2018

Joint Research Centre
Universitat Pompeu Fabra
Why use ML methods in criminal justice?

• Judge decisions are affected by extraneous factors
  [Danziger et al., 2011; Chen, 2016]

• Algorithms are not affected by cognitive bias

• There can be welfare gains: ML flight risk evaluation can yield substantial reductions in crime rate (with no change in jailing rate) or jailing rates (with no increase in crime rates)
  [Kleinberg et al., 2017]
Why **NOT** use ML methods in criminal justice?

- Machines can inherit human biases through biased data
  - [Barocas and Selbst, 2016]
- In many cases their outputs cannot be explained, so how can we justify?
- “They” can be racist
- There is a need for “fair” ML
Fairness in ML: the case of COMPAS

• ProPublica: COMPAS is unfair! [Angwin et al., 2016]

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted higher risk, did not reoffend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Predicted lower risk, did re-offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

• NorthPointe: COMPAS is fair!

Corbett-Davies et al., 2017
Fairness in ML: the case of COMPAS

Impossibility proofs: When base rates differ (in Broward County 51% vs. 39%), you cannot achieve calibration and equal FPR/FNR at the same time [Kleinberg et al., 2016; Chouldechova, 2017]

Also:
- No single threshold equalizes both FPR and FNR
  - Direct vs. indirect discrimination
- Imposing any fairness criterion has a cost in terms of public safety or defendants incarcerated
- Literature on fairML grows rapidly, but all based on US data

Corbett-Davies et al., 2017
What we do

• Look at European example: SAVRY in Catalonia

• We evaluate SAVRY against ML methods in terms of fairness and predictive performance

• We show some evidence that ML methods of risk assessments introduce unfairness and that their use in criminal justice should be fairness-aware
SAVRY

- Structured Assessment of Violence Risk in Youth (SAVRY)
- Structures Professional Judgement
- Also used to assess the risk of (not only violent) crimes upon release
- Used to inform decisions on interventions
- Sample: Catalonia, 4752 youths aged 12-18, 855 with SAVRY, committed crime between 2002-2010, released in 2010, recidivism by 2015
SAVRY ≠ COMPAS

- Detailed and transparent risk assessment
- Based on 6 protective factors
- Based on 24 risk factors: Historical, Social/Contextual, Individual
- We evaluate the sum of 24 risk factors (low, medium, high) against ML methods
Base rates differ

<table>
<thead>
<tr>
<th></th>
<th>Recidivated</th>
<th></th>
<th>Not Recidivated</th>
<th></th>
<th>Difference</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Diff</td>
<td>Std. Dev.</td>
</tr>
<tr>
<td>savry</td>
<td>0.20</td>
<td>(0.40)</td>
<td>0.17</td>
<td>(0.38)</td>
<td>-0.03**</td>
<td>(0.01)</td>
</tr>
<tr>
<td>female</td>
<td>0.1</td>
<td>(0.31)</td>
<td>0.2</td>
<td>(0.41)</td>
<td>0.1***</td>
<td>(0.0)</td>
</tr>
<tr>
<td>foreign</td>
<td>0.44</td>
<td>(0.50)</td>
<td>0.32</td>
<td>(0.47)</td>
<td>-0.12***</td>
<td>(0.01)</td>
</tr>
<tr>
<td>national group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>central/south</td>
<td>0.17</td>
<td>(0.38)</td>
<td>0.15</td>
<td>(0.35)</td>
<td>-0.03**</td>
<td>(0.01)</td>
</tr>
<tr>
<td>american</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU</td>
<td>0.05</td>
<td>(0.22)</td>
<td>0.05</td>
<td>(0.22)</td>
<td>0.00</td>
<td>(0.01)</td>
</tr>
<tr>
<td>magribian</td>
<td>0.2</td>
<td>(0.40)</td>
<td>0.1</td>
<td>(0.30)</td>
<td>-0.1***</td>
<td>(0.0)</td>
</tr>
<tr>
<td>age maincrime</td>
<td>15.59</td>
<td>(1.07)</td>
<td>15.92</td>
<td>(1.07)</td>
<td>0.33***</td>
<td>(0.03)</td>
</tr>
<tr>
<td>2 prior</td>
<td>0.21</td>
<td>(0.40)</td>
<td>0.21</td>
<td>(0.41)</td>
<td>0.00</td>
<td>(0.01)</td>
</tr>
<tr>
<td>3+ prior</td>
<td>0.10</td>
<td>(0.30)</td>
<td>0.09</td>
<td>(0.28)</td>
<td>-0.01</td>
<td>(0.01)</td>
</tr>
<tr>
<td>violent maincrime</td>
<td>0.52</td>
<td>(0.50)</td>
<td>0.54</td>
<td>(0.50)</td>
<td>0.02</td>
<td>(0.02)</td>
</tr>
<tr>
<td>action mediat+rep</td>
<td>0.28</td>
<td>(0.45)</td>
<td>0.33</td>
<td>(0.47)</td>
<td>0.05***</td>
<td>(0.01)</td>
</tr>
<tr>
<td>action execution</td>
<td>0.29</td>
<td>(0.45)</td>
<td>0.35</td>
<td>(0.48)</td>
<td>0.06***</td>
<td>(0.01)</td>
</tr>
<tr>
<td>action duration</td>
<td>118.67</td>
<td>(193.55)</td>
<td>141.05</td>
<td>(216.00)</td>
<td>22.38***</td>
<td>(6.38)</td>
</tr>
</tbody>
</table>

N = 1622 3130

Note: Authors’ calculations.
Performance

ROC Curve

- Logit on savry ML, AUC=0.66
- SAVRY overall score, AUC=0.64
Performance

ROC Curve

- logit on savry ML, AUC=0.66
- SAVRY overall score, AUC=0.64
- logit on demog. & crm. hist., AUC=0.69
Performance
Performance

![ROC Curve Graph]

- Logit on savry ML, AUC=0.66
- SAVRY overall score, AUC=0.64
- Logit on demog. & crim. hist., AUC=0.69
- Logit on demog. + savry, AUC=0.7
- Logit on demog. + more data, AUC=0.71

**False Positive Rate (Precision)** vs **True Positive Rate (Recall)**
Performance

Crime rate for different thresholds

- Logit on savry ML, AUC=0.66
- SAVRY overall score, AUC=0.64
- Logit on demog. + savry, AUC=0.7

Share of assigned interventions

Crime rate
Fairness

Error rate balance (Chouldechova 2017)

logit on savry ML, Estranger, AUCg=0.68
SAVRY overall score,Estranger, AUC=0.64
Fairness

Error rate balance (Chouldechova 2017)

- logit on savry ML, Stranger, AUCg=0.68
- SAVRY overall score, Stranger, AUC=0.64
- logit on demog. & crim. hist., Stranger, AUCg=0.7
- logit on demog. + savry, Stranger, AUCg=0.72

FPR Disparity

FNR Disparity
Fairness

Error rate balance (Chouldechova 2017)

- logit on savvy ML, Stranger, AUCg=0.68
- SAVRY overall score, Stranger, AUC=0.64
- logit on demog. & crim. hist., Stranger, AUCg=0.7
- logit on demog. + savvy, Stranger, AUCg=0.72
- logit on demog. + more data, Stranger, AUCg=0.72
Summary and Outline

- ML yields a more precise risk assessment
- When base rates differ, ML methods have to be fairness aware
- Use rich information:
  - for a transparent mitigation of unfairness
  - to adjust features that have a substantial effect on increasing unfairness
  - to refocus analysis away from tensions/tradeoffs towards better targeted interventions
- Further Analysis on human-algorithm interaction: RisCanvi
Thank you!

Any questions?
You can find me at songul.tolan@ec.europa.eu

Find HUMAINT at https://ec.europa.eu/jrc/communities/community/humaint
Find Carlos at http://chato.cl/